# Desafios na recolha de informação baseada na Wikipédia portuguesa com o Págico

João Miranda
Instituto Superior Técnico
joaocarvalhomiranda@ist.utl.pt

**Resumo**

O Págico foi uma iniciativa de recolha de informação em português, em que se usou uma cópia local da Wikipédia portuguesa para responder a 150 tópicos sobre temas referentes à lusofonia. As perguntas não tinham um número limitado de respostas previamente conhecido. O sistema de apoio ao Págico permitia navegar e pesquisar na cópia local da Wikipédia e apresentar as respostas e justificações aos tópicos. Este artigo sumariza os principais desafios encontrados e a metodologia usada com a participação humana nesta iniciativa.

**Palavras chave**

Avaliação conjunta, recolha de informação, Págico, lusofonia, Wikipédia

## 1 Introdução

O Págico foi organizado pela Linguateca e surgiu como sequência do GikiCLEF (Santos et al., 2010), uma iniciativa de recolha de informação em diferentes línguas. Ao contrário do GikiCLEF, o Págico foca-se apenas numa língua, o português, utilizando uma versão da Wikipédia portuguesa e perguntas em português a temas de cariz lusófono.

Um dos objectivos do Págico era avaliar e poder comparar o desempenho dos sistemas de resposta automática, e também humana, a perguntas para as quais não há um número limitado de respostas previamente conhecido. As respostas tinham de ser justificadas, o que significa que não bastava indicar as respostas, era necessário justificá-las. As respostas a dar eram os próprios artigos da Wikipédia correspondentes à resposta pretendida: a resposta tinha de ser ela própria uma entrada da Wikipédia e não páginas onde a resposta estivesse presente. Por exemplo, algumas das aves descritas na página *Aves de Angola* da cópia local da Wikipédia não tinham página própria criada e não podiam, por isso, ser dadas como resposta ao tópico *Aves de Angola*.

A colecção de avaliação disponibilizada pelo Págico tinha 150 tópicos para resposta. A cada tópico correspondia um, ou mais, dos seguintes temas: Artes, Ciência, Cultura, Desporto/esportes, Economia, Geografia, Letras, Política. Cada tópico tinha atribuído um ou mais dados geográficos que variavam entre: Angola, Brasil, CaboVerde, Geral, GuinéBissau, Lusofonia, Macau, Moçambique, Portugal, SãoToméPríncipe, Timor.

O Págico estava assente no sistema SIGA (Costa, Mota e Santos, 2012) que permitia visualizar os tópicos, navegar e pesquisar na versão local da Wikipédia, e apresentar as respostas aos tópicos e justificações seleccionadas. Foi utilizada uma cópia da Wikipédia portuguesa do dia 25 de Abril de 2011.

Na secção 2 deste artigo mencionam-se alguns trabalhos anteriores a esta iniciativa. Na secção 3 apresenta-se a motivação para a participação humana no Págico. Na secção 4 descreve-se a metodologia utilizada para responder aos tópicos. Na secção 5 referem-se os desafios encontrados. Na secção 6 apresentam-se algumas conclusões e breves sugestões de melhoria do sistema.

## 2 Trabalho relacionado

A Linguateca organizou anteriormente várias iniciativas de avaliação conjunta. A iniciativa predecessora do Págico, o GikiCLEF, foi organizada em 2009 no âmbito do CLEF[1]. O GikiCLEF disponibilizava 50 tópicos em 9 línguas europeias, para os quais não havia um número determinado de respostas conhecidas. O GikiCLEF surgiu na sequência do GikiP (Santos et al., 2009), organizado um ano antes e que continha apenas 15 tópicos em 3 línguas europeias.

[1] http://www.clef-initiative.eu

## 3 Motivação

A motivação para a participação humana no Págico compreendeu diferentes pontos. Por um lado, poder participar num desafio de avaliação conjunta é, por si só, interessante. Por outro lado, é estimulante pôr à prova as capacidades de recolha de informação num sistema de pesquisa limitado. Aprender coisas novas em matérias fracamente dominadas é, também, enriquecedor.

Outra das motivações foi responder ao desafio da luta homem-máquina que sempre despertou o interesse da Inteligência Artifical.

## 4 Metodologia

Na busca de informação as técnicas de pesquisa podem basear-se nos termos de partida ou nos termos de chegada. A diferença é subtil mas muito relevante e baseia-se na distinção entre aquilo por que queremos procurar e aquilo que queremos encontrar. Por exemplo, pesquisar por *aves de Angola* é diferente de pesquisar por *"é uma ave de Angola"*: a segunda opção permite obter resultados directos, se os houver.

Há quatro cenários principais na busca de informação que influenciam o sucesso de uma pesquisa:

1. sabemos onde está determinada informação e sabemos identificá-la;
2. sabemos onde está determinada informação mas não sabemos identificá-la;
3. não sabemos onde está determinada informação mas sabemos identificá-la;
4. não sabemos onde está determinada informação nem sabemos identificá-la.

É particularmente difícil encontrar informação quando não sabemos onde ela está nem sabemos identificá-la. É o caso de quando não se conhece a resposta a um tópico nem se vislumbram quais os artigos que nos poderão ajudar a respondê-lo. Por motivos de gestão do tempo de resposta, os tópicos que correspondiam a este cenário foram relegados para análise posterior, que não chegou a acontecer, em favor dos que se enquadravam nos três primeiros cenários.

Dos 40 tópicos respondidos, 27,5% enquadravam-se no primeiro cenário, 47,5% no segundo e 25% no terceiro. Por exemplo, o tópico *Línguas faladas em Timor Leste* enquadrava-se no primeiro cenário: conheciam-se duas respostas possíveis de antemão e bastava apenas verificar se existiam os artigos correspondentes na cópia local da Wikipédia.

Não houve uma ordenação intencional definida na escolha dos tópicos a responder. Na progressão das respostas às perguntas, para além de se preterirem as que correspondiam ao cenário 4, seguiram-se, em geral, as seguintes linhas de orientação:

1. se o tema de uma dada pergunta era familiar partia-se para a resposta pesquisando pelo artigo tido como provável de conter a resposta;
2. se o tema não era familiar, tentava-se uma pesquisa com uma expressão de busca contendo um dos termos ou expressões existentes na pergunta;
3. se uma pesquisa não oferecia, à segunda tentativa, resultados satisfatórios, partia-se para uma nova pergunta.

Como as perguntas eram de complexidade diferente, enquanto para algumas a resposta foi obtida navegando por poucos artigos, houve outras em que foi necessário consultar mais artigos para chegar às respostas e às justificações. Por exemplo, o tópico *Países que venceram a Copa do Mundo em uma disputa de pênaltis* foi respondido com 2 respostas e 2 justificações diferentes, enquanto o tópico *Escritores cabo-verdianos com obra publicada em crioulo* foi respondido com 4 respostas e nenhuma justificação. O grau de complexidade das perguntas era influenciado por diferentes factores, como o nível de familiaridade com o tema, o número de respostas a dar, o número de páginas que era necessário consultar e cruzar para responder a um tópico e justificar a resposta, ou o tempo dispendido até encontrar uma resposta considerada correcta. Dos 40 tópicos respondidos, 22,5% foram considerados fáceis, 27,5% de dificuldade média e 50% foram considerados difíceis.

Depois de respondidas, as respostas e justificações foram revistas para confirmação de correcção.

Para a pesquisa não foi dada particular atenção ao *Tema* e *Dados geográficos* de cada tópico. Foram usados, essencialmente, os termos, expressões e entidades mencionadas presentes em cada tópico.

## 5 Desafios encontrados

Houve diferentes desafios encontrados com a participação no Págico. Em primeiro lugar, foi necessária a familiarização com os termos habitualmente utilizados nesta área (ex.: *tópicos* e *corridas*).

Em segundo lugar, houve uma dificuldade inicial transitória em perceber o que devia ser apresentado como resposta e o que devia ser apresentado como justificação. O triângulo pergunta-resposta-justificação seguia um formato próprio e as respostas não eram dadas como num jogo de perguntas vulgar. As respostas eram os próprios artigos da cópia local da Wikipédia, e não artigos que pudessem conter a resposta, ou que a permitissem deduzir de forma indirecta. Nem todos os tópicos correspondiam a perguntas feitas de forma interrogativa. Enquanto umas eram interrogações directas (ex.: *Quem descobriu São Tomé e Príncipe?*) outras eram feitas de forma indirecta (ex.: *Empresários lusófonos com uma fortuna considerável*). Na Tabela 1 apresentam-se alguns exemplos de tópicos do Págico, e respectivos *Temas* e *Dados geográficos*.

As respostas não eram, em geral, directas, isto é, os tópicos foram construídos de forma a que fosse necessário relacionar artigos para encontrar uma resposta e poder justificá-la usando outros artigos, consoante as necessidades. A ideia era retirar o ênfase da extracção de respostas directamente a partir do texto e testar a capacidade de resposta quando é preciso cruzar informação de artigos diferentes.

As limitações do sistema de pesquisa foram uma das maiores dificuldades encontradas. O sistema de pesquisa apenas permitia procurar no título da página e pela ordem dos termos introduzidos na expressão de busca. Para responder às perguntas, ou se sabia de antemão a resposta à pergunta e se procurava o respectivo artigo, ou se iniciava a pesquisa com uma expressão de busca que se julgasse ser um bom ponto de partida para encontrar uma resposta. Como as respostas não eram fechadas, ou seja, não havia um número prévio limitado de respostas conhecidas, era por vezes difícil decidir se uma pergunta estava satisfatoriamente respondida e justificada; e se se deveria partir para outra pergunta ou, antes, procurar mais respostas e justificações para a pergunta corrente. Por exemplo, ao tópico *Instrumentos musicais de origem africana comuns no Brasil* deram-se 1 resposta e 1 justificação, enquanto ao tópico *Telenovelas brasileiras passadas no tempo da escravatura no Brasil* se deram 4 respostas e nenhuma justificação. Por seu turno, ao tópico *Aves de Angola* apresentaram-se 6 respostas e 1 justificação igual para todas as respostas. Seriam 6 respostas suficientes? Bastariam 3, ou 10 seria melhor?

A Wikipédia original pode ser usada como recurso de tradução: sabendo um termo numa língua de partida, é possível usar a correspondência de artigos entre línguas para encontrar o termo tido como equivalente numa língua de destino. O mesmo se passa ao procurar informação cruzada entre artigos de diferentes línguas, uma vez que os artigos têm, frequentemente, informação e completude variada entre elas. Por isso, ao pesquisar na Wikipédia é, muitas vezes, vantajoso partir de uma língua diferente para encontrar a informação pretendida. O cruzamento de artigos em mais do que uma língua permite uma abrangência maior de informação que falta na versão monolingue usada no Págico.

A Wikipédia oferece outras capacidades de cruzamento e extracção de informação: desde as hiperligações entre artigos até à categorização e hierarquia de artigos. As categorias são uma funcionalidade que permitiria resposta facilitada a muitos tópicos, mas tendo sido esvaziadas na versão utilizada no Págico, revelaram-se inúteis: as categorias existiam mas não continham informação. Não era, por isso, possível ir à categoria *Aves de Angola* para responder ao tópico *Aves de Angola*. Tornava-se, pois, mais difícil relacionar informação, o que seria feito com relativa facilidade se se pesquisasse a mesma informação na Wikipédia original.

De um ponto de vista mais geral, em qualquer matéria de estudo as tarefas são facilitadas se houver um fio condutor: desde a investigação criminal, aos processos de memorização cerebral, à comunicação entre diferentes unidades de uma empresa. Partir esta interligação de informação na versão local da Wikipédia é reduzir as capacidades de sucesso de resposta aos tópicos, uma vez que obriga a navegar e a ler os artigos de forma mais exaustiva. Significa, também, colocar carga adicional de processamento e tempo dispendido na pesquisa de páginas relacionadas com o tema que se procura. Do ponto de vista humano, este tempo pode não ser aceitável quando se pretende responder a uma pergunta de forma completa. Na verdade, se o objectivo era obrigar a avaliação humana a cingir-se à pesquisa de páginas isoladamente e a relacionar depois a informação entre si, o objectivo foi, dessa forma, atingido. Mas é um retrocesso na forma como nos habituámos a lidar com a informação e a organizá-la para tornar mais simples a pesquisa e navegação. O ser humano tem capacidades notáveis de relacionamento e cruzamento de informação mas tem limitações no que respeita à quantidade de dados que pode analisar ao mesmo tempo. Para um humano, é difícil gerir muita informação em simultâneo. E esse é o impacto mais visível que os computadores introduziram em diversos campos,

| | Tópicos | Temas | Dados geográficos |
|---|---|---|---|
| Pagico_008 | Telenovelas brasileiras passadas no tempo da escravatura no Brasil | Artes, Letras | Brasil |
| Pagico_027 | Doenças letais comuns em países lusófonos transmitidas por mosquitos | Ciência | Lusofonia |
| Pagico_098 | Cidades dos Estados Unidos que tiveram forte imigração portuguesa | Geografia, Letras | Portugal |
| Pagico_119 | Pratos típicos da gastronomia de Cabo Verde | Cultura | Cabo Verde |
| Pagico_135 | Aves de Angola | Ciência, Geografia | Angola |

Tabela 1: Exemplo de tópicos no Págico

incluindo o da Linguística Computacional.

Entre os pontos favoráveis do processo de resposta incluíam-se a ausência de tempo limite para responder a uma pergunta e a não contabilização desse tempo para efeitos de avaliação de desempenho.

Por último, o elevado número de perguntas fez antever que seria difícil responder a todos os tópicos em pouco tempo. Este foi o principal factor que levou a que não tenham sido todos respondidos. Preferiu-se, também, procurar responder a um número mais reduzido de tópicos mas com uma maior completude de respostas e justificações julgadas correctas.

## 6 Conclusões

As ideias principais a reter com a participação nesta iniciativa estão relacionadas com as principais dificuldades sentidas:

1. as categorias tinham sido esvaziadas, o que inibia a resposta imediata a perguntas em que a informação podia ser extraída directamente das páginas de categoria;

2. mais importante, isso dificultava a navegação contínua entre páginas com pontos comuns entre si, uma vez que as categorias também servem para facilitar a agregação de páginas relacionadas, e isso é uma boa ajuda para extrair informação (que não constando desta versão modificada da Wikipédia, consta da versão original).

Humanamente, é difícil adivinhar que páginas conterão determinada resposta nos casos em que não se domina a matéria analisada, e isso notou-se muito ao procurar a informação usando a interface de pesquisa disponibilizada no Págico. Essa função é facilitada pelos motores de busca, e é a diferença óbvia entre:

1. pesquisar qualquer informação numa enciclopédia em papel, onde é preciso saber em que entradas se irá procurar o que pretendemos, ou seja, é preciso saber por onde começar, ou:

2. pesquisar num motor de busca avançado, como o Google, onde esse requisito não é importante.

Na primeira é necessário prever as entradas, e parte-se das entradas para os conteúdos; na segunda, faz-se o caminho inverso e descobrem-se facilmente as entradas a partir dos conteúdos.

Sentiu-se que seria mais fácil usar um motor de busca avançado para descobrir que páginas da Wikipédia conteriam determinada informação do que vaguear ao acaso entre artigos na esperança de encontrar as páginas que respondessem ao que se precisava. Da mesma forma que numa enciclopédia em papel seria necessário percorrer os artigos que se considerasse levarem à informação pretendida, também aqui na versão da Wikipédia do Págico foi sentida essa imposição. Mesmo que a informação estivesse lá, poderia não ser descoberta por não se saber onde procurá-la.

Do ponto de vista da luta homem-máquina, é aí que um sistema avançado de recuperação de informação ganha a um humano: o sistema encontra a informação em segundos e é capaz de procurar em toda a enciclopédia num instante, e um humano não é capaz de o fazer. Enquanto um humano anda de artigo em artigo a descobrir que missionários estiveram no Brasil no tempo dos Descobrimentos, um bom motor de busca faz isso num instante: tem as páginas que falam de missionários todas indexadas, todas em "memória", e o humano não.

Mas há uma coisa em que os humanos são melhores do que um motor de busca avançado: podem jogar com as expressões de busca e conduzir a pesquisa como querem, e um motor de busca não saberia nunca fazer isso. Em última instância são os humanos que verificam se os resultados que ele dá correspondem ao que procuram e contêm a resposta pretendida. São os humanos que refinam as pesquisas se entenderem que os resultados

não correspondem às expectativas. Um motor de busca avançado sabe que os resultados que dá têm aquilo por que se procurou, mas não sabe se o utilizador humano encontrou aquilo que de facto procurava.

No geral, o que se sentiu foi uma regressão nas capacidades de pesquisa. Encontrou-se extrema dificuldade em encontrar informação que se encontraria de forma fácil com outros instrumentos. Talvez seja reflexo da habituação às facilidades que os motores de busca vieram trazer.

Há três pontos-chave que permitiriam responder com maior sucesso e em menos tempo aos tópicos do Págico:

1. as categorias da Wikipédia;
2. os conteúdos noutras línguas;
3. um sistema de pesquisa que permitisse pesquisar em todo o artigo e não apenas no título.

Estes três pontos reflectem a diferença entre conseguir responder aos tópicos usando apenas o sistema disponibilizado, um sistema controlado do ponto de vista de validação, ou usando as ferramentas livremente disponibilizadas na Internet, mas que não permitiriam igualdade de recursos utilizados do ponto de vista de avaliação conjunta.

A interface disponibilizada pelo SIGA, apesar de simples, é suficiente para aquilo a que se propõe.

## Agradecimentos

Agradeço à equipa do Págico a oportunidade de ter participado nesta iniciativa e, em especial, à Cristina Mota pelo acompanhamento e ajuda ao longo do tempo.

Agradeço a apreciação e pertinência das observações de Cláudia Freitas e Stella Tagnin que ajudaram a tornar este artigo mais esclarecedor e interessante.

## Referências

Costa, Luís, Cristina Mota, e Diana Santos. 2012. SIGA, a Management System to Support the Organization of Information Retrieval Evaluations. Em Helena Caseli, Aline Villavicêncio, António Teixeira, e Fernando Perdigão, editores, *Computational Processing of the Portuguese Language, PROPOR'2012*, pp. 284–290, Berlim/Heidelberg. Springer.

Santos, Diana, Luís Miguel Cabral, Corina Forascu, Pamela Forner, Fredric Gey, Katrin Lamm, Thomas Mandl, Petya Osenova, Anselmo Peñas, Álvaro Rodrigo, Julia Schulz, Yvonne Skalban, e Erik Tjong Kim Sang. 2010. Gikiclef: Crosscultural issues in multilingual information access. Em Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, e Daniel Tapias, editores, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may, 2010. European Language Resources Association (ELRA).

Santos, Diana, Nuno Cardoso, Paula Carvalho, Iustin Dornescu, Sven Hartrumpf, Johannes Leveling, e Yvonne Skalban. 2009. GikiP at GeoCLEF 2008: Joining GIR and QA forces for querying Wikipedia. Em Carol Peters, Tomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J.F.Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, e Viviane Petras, editores, *Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*. Springer, pp. 894–905.